Predicting flu cases using dynamic regression with Google Flu Trend

Mohammad Khan, MSc; Sangook Kim, MSc; Kuan Liu, MMath; Michael Moon, BASc
Department of Biostatistics, Dalla Lana School of Public Health, University of Toronto

UNIVERSITY OF TORONTO
DALLA LANA SCHOOL OF PUBLIC HEALTH

Abstract

Background

Early prediction of influenza outbreak can significantly reduce its burden and impact. Google uses a validated algorithm to provide a region-specific estimates (GFT) of influenza activities in real-time.

Objective

The first objective was to evaluate the association between GFT and various respiratory illness cases in Canada. The second objective was to predict the peak in the number of influenza cases.

Method

Non-parametric Spearman's correlation and and cross-correlations were used to assess the association between GFT and respiratory illness outbreaks in Canada. A number of seasonal ARIMA and dynamic regression models with GFT were applied to forecast the trend of influenza cases. Model comparisons were conducted using cross-validation.

Results and Conclusion

The results suggested that GFT was significantly associated with the number of influenza tests. Furthermore, GFT improved the three-week forecast of the timing of influenza outbreaks.

References

Google Flu Trends. Retrieved May 2, 2016, from https://www.google.org/flutrends/about/.

Grothendieck, G. (2012). dyn: Time Series Regression. R package version 0.2-9. https://CRAN.R-project.org/package=dyn

Martin, L. J., Lee, B. E., & Yasui, Y. (2015). Google Flu Trends in Canada: A comparison of digital disease surveillance data with physician consultations and respiratory virus surveillance data, 2010–2014. Epidemiol. Infect. Epidemiology and Infection, 144(02), 325-332.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Shumway, R. H., & Stoffer, D. S. (2005). Time series analysis and its applications: With R examples. New York: Springer.

Thompson, L. H., Malik, M. T., Gumel, A., Strome, T., & Mahmud, S. M. (2014). Emergency department and 'Google flu trends' data as syndromic surveillance indicators for seasonal influenza. Epidemiol. Infect. Epidemiology and Infection, 142(11), 2397-2405.

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag.

Zeileis, A. (2014). dynlm: Dynamic Linear Regression. R package version 0.3-3. URL http://CRAN.R-project.org/package=dynlm.

Background

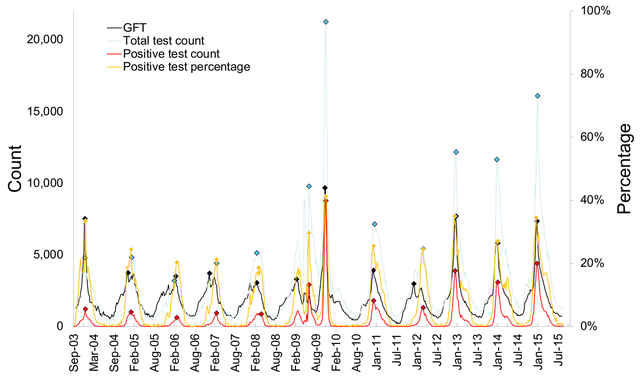
- **FluWatch:** Canada's national surveillance system collecting weekly counts of all tests performed and positive tests for influenza and influenza-like illnesses including parainfluenza virus, adenovirus, human metapneumovirus, rhinovirus and coronavirus
- **Google Flu Trend (GFT):** Influenza case estimates using Google search keywords provided by Google
- This research aims to investigate the association between GFT with FluWatch data and whether GFT can help predicting the outbreak curve

Objective

- To identify the strength and type of the association between:
  - GFT and FluWatch's total and positive test counts for influenza in Canada
  - GFT and FluWatch's total and positive test counts for other respiratory viruses in Canada
- To predict the timing of influenza outbreaks

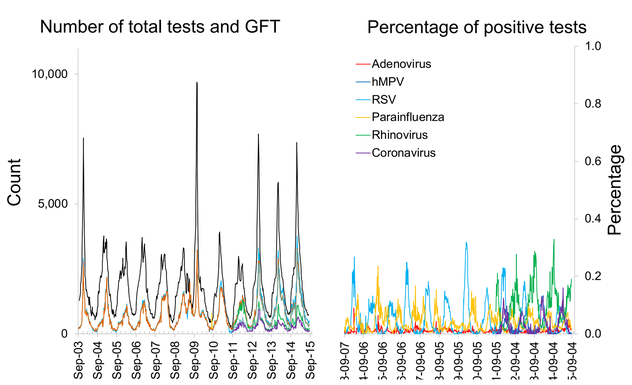
Data

Figure 1. Weekly data of GFT and FluWatch influenza cases in Canada.



- ◆ represent the maximum annual counts/percentages
- H1N1 pandemic wave between Oct and Dec 2009

Figure 2. Weekly data of GFT and FluWatch other influenza-like cases in Canada.



- GFT estimates follow the trend of the total test counts for both influenza and other respiratory illnesses
- Percentage and positive test counts less discernible with GFT

Methods

Association

1. **Strength of correlation:** Assess overall correlation between GFT and FluWatch using Spearman's correlation
2. **Time-series correlation:** Identify the number of lags or leads with the strongest cross-correlation between GFT and FluWatch

Prediction

1. **Time-series model:** Constructed Seasonal ARIMA models to forecast a number of weeks ahead

$$\text{ARIMA}(2,1,0)\times(1,1,0)_{52}$$

2. **Dynamic regression model:** Constructed regression models with leads with and without GFT

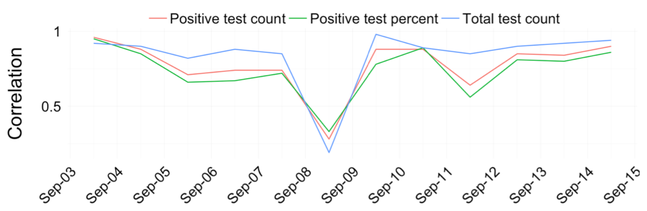
$$\text{FLU}_t = \alpha + \beta_1 \text{FLU}_{t-l_F} + \beta_2 \text{GFT}^*_{t-l_G} + \beta_3 \text{FLU}_{t-52} + \epsilon_t$$

- FLU_t is the positive influenza test count at time t
- GFT_t is the GFT estimate at time t
- models constructed with and without $\beta_2 \text{GFT}^*_{t-l_g}$

3. **Model selection:** Compared 5-to-1 year cross-validation mean squared prediction errors for model selection

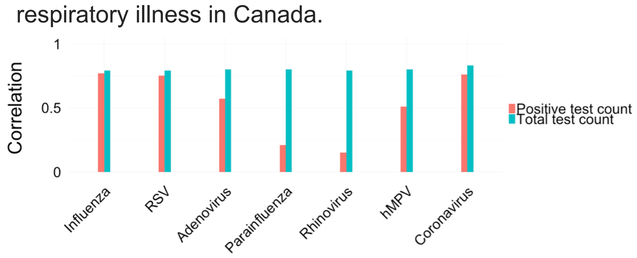
Association

Figure 3. Annual Spearman's correlations between GFT and influenza cases in Canada.



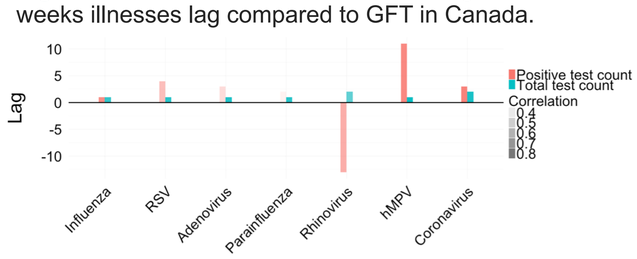
- GFT highly correlated with all three series
- Low correlation in 2008-2009 during H1N1 pandemic

Figure 4. Overall Spearman's correlations for each respiratory illness in Canada.



- Higher correlation with the total test count across all illnesses

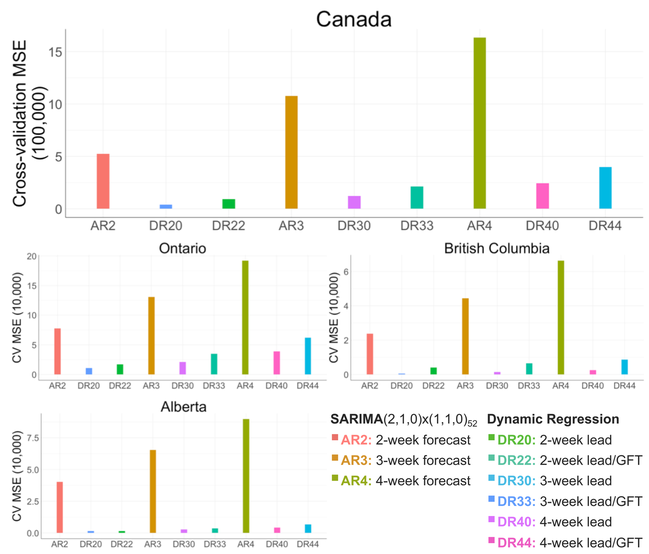
Figure 5. Cross-correlation results showing number of weeks illnesses lag compared to GFT in Canada.



- GFT leads FluWatch in general
- For influenza, 1 week lead had the strongest correlation

Prediction

Figure 6. Cross-validation mean square errors for model selection.



- Selected 3-week forecasting models based on the cross-validation MSE while maximizing forecast leads

Figure 7. 3-week forecasts from the Seasonal ARIMA model between Dec 2012 to Aug 2015.

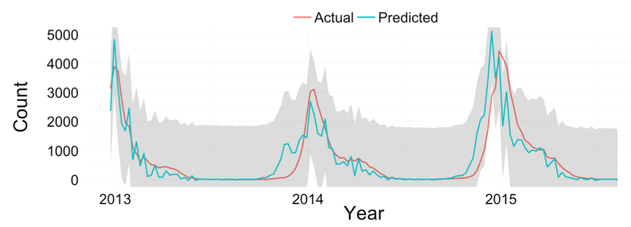
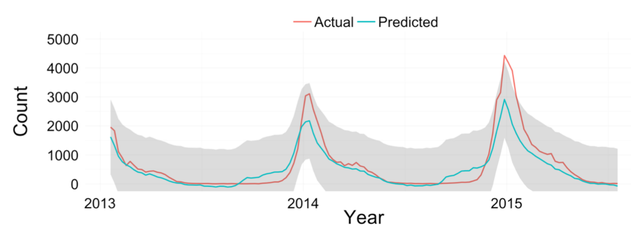


Figure 8. Prediction from the dynamic regression model with 3-week lead and GFT model between Dec 2012 to Aug 2015.



- MSE for the forecasts from the selected dynamic regression model was approx. 210,000 compared to 1,080,000 for the selected SARIMA

Discussion

- GFT most associated with the total number of tests for respiratory illnesses
- GFT leads by 1 week in general
- GFT is not a good indicator for pandemic outbreaks
- Dynamic regression model provides smoother prediction for better identification of annual peaks
- Including GFT in the dynamic regression model improved prediction performance
- Limited to 3-week forecasts and predicting local peaks
- Limited prediction accuracy for the count of positive influenza cases